Data Intake Report — G2M Cab Investment Case

Prepared by: Jenisa Thapa
Date: April 18, 2025

---

1. Overview of Datasets:

- **Cab_Data.csv**: Contains details of cab rides including transaction ID, date of travel (in Excel date format), company (Pink or Yellow), city, kilometers traveled, price charged, and cost of trip.
- **Customer_ID.csv**: Contains demographic details for customers, including customer ID, gender, age, and monthly income.
- **Transaction_ID.csv**: Maps each transaction to a customer and indicates the payment mode (Cash or Card).
- **City.csv**: Lists U.S. cities, their populations, and total number of cab users per city.

---

2. Data Cleaning Performed:

- Converted Date of Travel in Cab_Data.csv from Excel numeric format to standard datetime.
- Converted Population and Users columns in City.csv from strings with commas to numeric (integers).
- Verified that there were **no missing values or duplicate records** in any of the datasets.
- Created a new column Month to analyze seasonal trends in cab usage.

---

3. Merging Strategy:

- Merged Cab_Data and Transaction_ID using Transaction ID to add customer and payment details.
- Then merged with Customer_ID using Customer ID to include demographic data.
- Finally, merged with City.csv on the City field to include city-level data like population and total users.
- Used a **left join** in the final merge to retain all cab ride records, even if city data was missing.

---

4. Final Notes:

- The merged dataset covers the time period from **January 2016 to December 2018**.
- This final master dataset was used to explore profit, revenue, ride counts, and customer behavior by company, location, and time.
- All analysis and visualizations were performed using Python (Pandas, Seaborn, and Matplotlib) in a Jupyter/Colab notebook.